bio-protocol Preprint

Not Peer Reviewed
Ask a question online

AcrIIA11 homolog discovery and phylogenetic placement

Kevin J Forsberg

Updated date: May 21, 2020

An abbreviated version of this protocol was published in eLIFE in Sep 2019

Functional metagenomics-guided discovery of potent Cas9 inhibitors in the human microbiome

DOI: 10.7554/eLife.46540

Detailed protocol

We thank the reader for an inquiry into this methods section. Hopefully the detailed protocol below, along with the original methods section, will help the reader reproduce these analyses for AcrIIA11 or any additional protein of interest. In some places, the software or databases that we used have been updated, so we provide version information where appropriate. We believe this protocol to be thorough but recognize that it encompasses many related analyses and will happily field any additional questions that this document does not fully address. As our manuscript was focused on anti-CRISPR proteins (Acrs), these methods are tailored to that purpose. However, they should also be generally applicable to other query proteins, but individual parameters may need to be optimized for different queries.

1) Blast a query Acr (e.g. AcrIIA11, QEH00205.1) against NCBI's 'nr' database using NCBI's web interface
2) Retrieve Blast hits with ≥35% amino acid identity and that cover ≥75% of the query
    a. These proteins were included in the phylogenetic trees shown in Figure 4B and Figure 4—figure supplement 4A, along with additional homologs retrieved via a blastP query against the IMG/VR database. All sequences used in these figures are listed in supplementary table S8.
    b. The IMG/VR homologs were retrieved by blasting against the Jan. 1, 2018 database release with an e-value cutoff of $1e^{-10}$. The database is available here: https://genome.jgi.doe.gov/portal/IMG_VR/IMG_VR.download.html
    c. To create the final phylogenetic trees, a subset of IMG_VR proteins hits were selected to maximally sample phylogenetic space. These proteins are listed in supplementary table S8.
3) To phylogenetically classify the parent genome for each Acr homolog, the source genome for each NCBI blastP hit was downloaded in an automated manner using NCBI's E-utilities functions (https://www.ncbi.nlm.nih.gov/books/NBK25500/).
    a. If the user is uncomfortable with the e-utilities functions, this can also be done individually for each hit, via the web browser as follows. For the NCBI protein entry that corresponds to each hit:
        i. View the Identical Protein Report, and then
        ii. Click the link for the 'CDS Region in Nucleotide', and then
        iii. On the right panel, click on the 'assembly' link under the heading 'Related Information', and then
        iv. Download the raw fasta file for this genome assembly
4) Each assembled genome fasta file can be assigned to a phylogeny using the 'classify_wf' workflow in the GTDB toolkit. We used v0.1.3 of this software, which was the latest version at the time of analysis. As of now, the latest version can be found at https://github.com/Ecogenomics/GTDBTk.
    a. Genomes were classified using default parameters, per the GTDB toolkit documentation.
    b. This paper describes the GTDB toolkit: Chaumeil PA, et al. 2019. GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, btz848.
5) To visualize a protein's distribution across the bacterial tree of life, we used AnnoTree (http://annotree.uwaterloo.ca/).
    a. AnnoTree is a web tool for visualization of genome annotations across large phylogenetic trees and uses same the GTDB classification scheme used by the GTDB toolkit.
        i. Described further here: Mendler et al. 2019 *Nucleic Acids Research*, Volume 47, Issue 9, 21 May 2019, Pages 4442–4448, https://doi.org/10.1093/nar/gkz246
    b. At the time of analysis, we used AnnoTree v1.0. Via the web-browser, it is possible to zoom to various phylogenetic depths and download the newick trees or the vector-art for the displayed tree. It is also possible to highlight particular taxonomies or predicted protein functions on the tree of life. These functions were used to help create Figures 4A and Figure 4 – figure supplement 2, with additional annotations added in Adobe Illustrator. The Cas9 and PFAM searches across bacterial genomes were performed using these web-browser capabilities. The resulting tab-delimited datasets were used for count data and statistical analyses, per the published methods section.

How to cite: (Readers should cite both the Bio-protocol preprint and the original research article where this protocol was used)

1. Forsberg, K. (2020). AcrIIA11 homolog discovery and phylogenetic placement. Bio-protocol Preprint. bio-protocol.org/prep321.
2. Forsberg, K. J., Bhatt, I. V., Schmidtke, D. T., Javanmardi, K., Dillard, K. E., Stoddard, B. L., Finkelstein, I. J., Kaiser, B. K. and Malik, H. S. (2019). Functional metagenomics-guided discovery of potent Cas9 inhibitors in the human microbiome. eLIFE. DOI: 10.7554/eLife.46540